DM n°8 – Modèles multivariés

Tutorat 2020-2021 : **12 QRUS**

Voilà un petit DM sur des QRUs du cours « Introduction aux modèles multivariés ». Je sais que ce cours est compliqué, et vraiment pas facile à comprendre, donc profitez de ce DM pour éclairer vos lanternes. Faites-le sérieusement, et bossez la correction de ouf ! Il n'y a pas beaucoup de QRUs dispos sur ce cours (vu qu'il est nouveau) donc n'hésitez pas à limite le refaire 2 fois (plutôt qu'une) ou à bien noter vos erreurs pour être au taquet. Je ne vous garantis pas que ces QRUs sont représentatifs de ceux que le prof peut faire parce que j'en ai aucune idée, mais ce DM reprend des points essentiels à la compréhension du cours, et sur lesquels le prof a l'air de vouloir vous interroger (cf Réponses du prof) et qui sont donc, je pense, à maîtriser +++ J'ai fait de mon mieux, j'espère que ça vous conviendra ! Des bisous et bon courage !

<u>QRU 1</u> : A propos des modèles multivariés, indiquez la proposition exacte :

A) Lorsque la variable à expliquer est binaire et non censurée, le modèle statistique adapté est la régression linéaire
B) Lorsqu'il n'existe qu'une seule variable explicative et qu'elle est qualitative, la régression linéaire donne un résultat proche d'un test du Khi-2
C) Si on souhaite expliquer la taille d'un enfant en fonction de la taille de sa mère, Y est la taille de l'enfant et X la taille de la mère
D) Les tests statistiques classiques (analyse univariée) prennent en compte les potentiels facteurs de confusion
E) Les propositions A, B, C et D sont fausses

<u>QRU 2</u> : A propos des modèles multivariés, indiquez la proposition exacte :

A) L'analyse multivariée permet de tester chacun des facteurs Y pouvant avoir une influence sur la variable X et de leur donner un coefficient
B) Le risque de premières espèce correspond au risque qu'on prend a priori de conclure à tort qu'un coefficient au moins aussi élevé soit dû au hasard
C) La p-value est calculée a posteriori et correspond à la probabilité qu'on a d'observer un coefficient au moins aussi élevé uniquement en raison du hasard
D) Lorsque p est inférieur au risque alpha, on accepte l'hypothèse nulle de nullité du coefficient
E) Les propositions A, B, C et D sont fausses

<u>QRU 3</u> : On souhaite savoir si le poids de l'enfant à la naissance était corrélé à l'âge de la mère (Age madame), au sexe de l'enfant, au rang de la grossesse et au fait qu'il ait une malformation. Indiquez la proposition exacte :

		Estimation [IC]	p	p global
Age madame		4.45 [-0.152, 9.0]	0.058	
Sexe	M vs F	138 [100, 180]	**<0.001**	
Rang grossesse	gemellaire vs unique	-285 [-335, -234]	**<0.001**	***<0.001***
	triple vs unique	-442 [-589, -295]	**<0.001**	
Malformation	oui vs non	-71.4 [-138, -4.87]	**0.035**	

A) On peut conclure que le poids de la mère influence le poids de l'enfant
B) Le fait d'être un garçon augmente significativement le poids de l'enfant
C) Globalement, avoir une grossesse multiple n'influence pas significativement le poids de l'enfant
D) Avoir une malformation augmente significativement le poids de l'enfant
E) Les propositions A, B, C et D sont fausses

<u>QRU 4</u> : A propos des modèles multivariés, indiquez la proposition exacte :

A) Si on souhaite expliquer la probabilité de naître du sexe masculin en fonction du régime alimentaire, on se trouve dans un problème de régression logistique
B) Lorsqu'il n'existe qu'une seule variable explicative et que celle-ci est qualitative, la régression logistique donne un résultat similaire à un test t de Student
C) On utilise la fonction logit dans le modèle de régression linéaire
D) L'odds ratio (ou OR), utilisé en régression linéaire, exprime force du lien entre X et Y, c'est le rapport des cotes
E) Les propositions A, B, C et D sont fausses

<u>QRU 5</u> : On souhaite savoir si le sexe de l'enfant (0 = féminin) dépend de la technique utilisée et de l'âge des parents.

		Odds Ratio [IC]	p	p global
Age madame		1.00 [0.974, 1.03]	0.9	
Age monsieur		0.995 [0.977, 1.01]	0.6	
Technique	ICSI vs FIV	0.836 [0.66, 1.06]	0.1	*0.2*
	IMSI vs FIV	0.836 [0.64, 1.09]	0.2	

Indiquez la proposition exacte :

A) On peut conclure que l'âge de la mère ou l'âge du père ont une influence sur la probabilité que l'enfant soit de sexe masculin
B) La technique ICSI modifie statistiquement la probabilité que l'enfant soit de sexe masculin par rapport à la FIV
C) La technique IMSI modifie statistiquement la probabilité que l'enfant soit de sexe masculin par rapport à la FIV
D) La technique ICSI modifie statistiquement la probabilité que l'enfant soit de sexe masculin par rapport à la technique IMSI
E) Les propositions A, B, C et D sont fausses

<u>QRU 6</u> : A propos des modèles multivariés, indiquez la proposition exacte :

A) Les analyses univariées permettent de prendre en compte les variables d'ajustement
B) Les analyses multivariées ne sont pas recommandées lorsqu'on cherche à établir un lien statistique entre plusieurs variables
C) La p-value quantifie le lien entre deux variables, tandis que les mesures d'association renseignent sur la significativité statistique
D) Les modèles statistiques multivariés permettent de mesurer à quel point un facteur agit sur la variable d'intérêt
E) Les propositions A, B, C et D sont fausses

<u>QRU 7</u> : Les notes à l'épreuve de première session d'anglais et de biostatistique de 60 étudiants inscrits en master en 2009 ont été analysées. Les statistiques descriptives résumées figurent dans le tableau suivant. Existe-t-il une relation entre la note d'anglais et la note de biostatistique en master ? Indiquez la proposition exacte :

	Anglais	Biostatistique
moyenne (m)	**13,2**	**12,7**
écart-type (s)	**1,5**	**2,6**
somme (anglais*biostat)	**10173,0**	

A) Il s'agit d'un problème de régression
B) On note l'hypothèse nulle : « H0 = Il existe une liaison entre la note d'anglais et la note de biostatistique chez les étudiants de master »
C) Pour pouvoir utiliser un test de corrélation ici, il est nécessaire d'observer une indépendance des observations, et une liaison linéaire entre les deux variables
D) En obtenant un paramètre calculé de 0,5, on conclut que les notes de 1ère session d'anglais et de biostatistique ne sont pas corrélées chez les étudiants de master, au risque alpha = 5%
E) Les propositions A, B, C et D sont fausses

<u>Énoncé des QRUs 8 à 12</u> : Une étude a été conduite sur un échantillon de 30 sujets pour déterminer si la valeur de la pression artérielle systolique dépendait de l'âge. Les statistiques descriptives sont présentées dans le tableau suivant.

	Age (an)	PAS (mmHg)
moyenne (m)	**45**	**143**
écart-type (s)	**15**	**23**
somme (âge*PAS)	**199576**	

<u>QRU 8</u> : Les conditions d'application à vérifier avant d'estimer les paramètres (pente et ordonnée à l'origine) de la droite de régression linéaire de la pression artérielle systolique en fonction de l'âge sont (indiquez LA proposition exacte) :

A) Un degré de signification p<0,05
B) Des effectifs théoriques attendus sous l'hypothèse nulle H0 tous supérieurs ou égaux à 5
C) Une dépendance entre les observations
D) Une liaison linéaire entre la pression artérielle systolique et l'âge
E) Les propositions A, B, C et D sont fausses

<u>QRU 9</u> : Dans la droite de régression de la pression artérielle systolique en fonction de l'âge (dont l'équation est PAS = α + β x âge), indiquez la proposition exacte :

A) L'âge est la variable dépendante
B) La pression artérielle systolique est la variable explicative
C) La pression artérielle systolique est la variable indépendante
D) L'âge est la variable expliquée
E) Les propositions A, B, C et D sont fausses

<u>QRU 10</u> : L'estimation du coefficient de la pente (b) de la droite de régression est de 1.0 et l'estimation de son écart-type (sb) est de 0.2. La valeur observée du test de la pente de la droite de régression est égale à (indiquez la proposition exacte) :

A) 2,048
B) 5
C) 0,05
D) 28
E) Les propositions A, B, C et D sont fausses

<u>QRU 11</u> : Le degré de signification (P-value) associé au test du coefficient de la pente de la droite de régression est inférieur à 0.001. Comment interpréter cette information ? Indiquez la proposition exacte :

A) La pente de la droite de régression diffère significativement de 0
B) La pente de la droite de régression est égale à 0
C) La pression artérielle systolique moyenne diffère significativement de l'âge moyen
D) La pente de la droite de régression est significativement inférieure à 0.001
E) Les propositions A, B, C et D sont fausses

<u>QRU 12</u> : A propos de l'estimation du coefficient de l'ordonnée à l'origine (a) de la droite de régression, indiquez la proposition exacte :

A) Une particularité de la droite de régression est de passer par le point moyen théorique de coordonnée (mx, my)
B) L'estimateur de l'ordonnée à l'origine b est déduit de la pente a et des coordonnées du point moyen (mx, my) : b=my –amx
C) On note a=my – bmx, avec mx = mPAS = 143 et my = mâge = 45
D) L'estimation du coefficient de l'ordonnée à l'origine (a) de la droite de régression est égale à 48
E) Les propositions A, B, C et D sont fausses

CORRECTION :

QRU 1 : C

A) Faux : on utilise la régression linéaire principalement pour des variables à expliquer qui sont quantitatives continues
B) Faux : résultat proche d'un test t de Student ++
C) Vrai
D) Faux : ils ne les prennent pas en compte, or ceux-ci sont fréquents en médecine, et il est donc nécessaire d'avoir recours à des méthodes statistiques plus complexes = les modèles statistiques de régression (analyse multivariée)
E) Faux

*Pour mieux comprendre : **FACTEURS DE CONFUSION** :*
Imaginons que l'on souhaite si les buveurs de café ont un risque plus élevé de développer un cancer du poumon. Si on fait un simple test statistique, on s'apercevra qu'il existe une association significative entre les deux. Or, dans ce cas, ne pas ajuster serait une erreur, car il est nécessaire de prendre en compte (entre autres) le tabagisme comme variable de confusion.
L'association significative trouvée par le test serait due à la fois à l'association statistique entre tabagisme et cancer, et à la fréquence de consommation du café plus fréquente chez les fumeurs, constituant donc un fameux biais de confusion.

QRU 2 : C

A) Faux : chacun des facteurs **X** pouvant avoir une influence sur la variable **Y**
B) Faux : ne soit PAS du au hasard
C) Vrai +++
D) Faux : on rejette l'hypothèse nulle
E) Faux

QRU3 : B

A) Faux : L'âge de la mère n'influence pas le poids de l'enfant ($p > 0.05$); pour chaque année supplémentaire, le poids de l'enfant augmente de 4.45g, avec un intervalle de confiance comprenant 0 : [-0.152, 9.0]
B) Vrai : +138g [100, 180]
C) Faux : avoir une grossesse multiple a pour conséquence un poids plus faible chez l'enfant (p global <0.001)
D) Faux : Avoir une malformation diminue significativement le poids de l'enfant (-71.4g [-138, -4.87])
E) Faux

QRU 4 : A

A) Vrai : on note Y est le sexe masculin, et X le régime alimentaire. La variable à expliquer est bien binaire
B) Faux : à un test du Khi-2
C) Faux : dans le modèle de régression logistique
D) Faux : utilisé en régression logistique
E) Faux

QRU 5 : E

A) Faux : Ni l'âge de la mère, ni l'âge du père n'ont une influence sur la probabilité que l'enfant soit de sexe masculin ($p > 0.05$)
B) Faux : La technique ICSI ne modifie pas statistiquement la probabilité que l'enfant soit de sexe masculin ($p > 0.05$) par rapport à la FIV
C) Faux : La technique IMSI ne modifie pas statistiquement la probabilité que l'enfant soit de sexe masculin ($p > 0.05$) par rapport à la FIV
D) Faux : Les deux techniques ICSI et IMSI ne sont pas comparées entre elles
E) Faux

QRU 6 : D

A) Faux : non, ce sont les analyses multivariées qui les prennent en compte
B) Faux : si, justement
C) Faux : c'est l'inverse
D) Vrai
E) Faux

QRU 7 : C

A) Faux : il s'agit d'un problème de corrélation. Il est possible que les 2 variables soient liées mais l'une n'est pas susceptible de dépendre de l'autre : il ne s'agit donc pas d'un problème de régression**.**
B) Faux : H0 = il n'existe PAS de liaison linéaire entre la note d'anglais et la note de biostatistique chez les étudiants de master
C) Vrai
D) Faux : on a r=0,5 > 0 donc on a une liaison positive : les notes de 1ère session d'anglais et de biostatistique sont positivement corrélées chez les étudiants de master
E) Faux

QRU 8 : D

A) Faux : Le degré de signification est déterminé a posteriori (cad après avoir calculé la valeur du test). Ce n'est pas une condition d'application du test qui doit être vérifiée a priori (i.e., avant de calculer la valeur du test)
B) Faux : (il s'agit d'une condition d'application du test du Chi2, *mais peu importe, ne retenez pas ça*)
C) Faux : une indépendance des observations
D) Vrai : le plus souvent vérifiée empiriquement (sur les données de l'échantillon) par l'examen du nuage de points
E) Faux

QRU 9 : E

A) Faux : La pression artérielle systolique est la variable dépendante
B) Faux : L'âge est la variable explicative
C) Faux : L'âge est la variable indépendante
D) Faux : La pression artérielle systolique est la variable expliquée
E) Vrai : on a : **X = l'âge est la variable explicative (= indépendante)**
et **Y = la pression artérielle systolique est la variable dépendante (= expliquée ou « à expliquer »)**

QRU 10 : B

A) Faux : (il s'agit de la valeur de tα pour 28 ddl)
B) Vrai : on effectue le test de la pente de la droite de régression -> $\frac{b}{s_b} \rightarrow t_{(n-2)ddl}$
C) Faux : (il s'agit de la valeur du risque alpha)
D) Faux : (il s'agit du nombre de degré de liberté test de la pente de la droite de régression pour un échantillon de 30 sujets**)** $t_o = \frac{1}{0,2} = 5$
E) Faux

QRU 11 : A

Pour résoudre ce QRU, on va :

1. Formuler les hypothèses du test de la pente de la droite de régression :
- H0 : la pente de la droite de régression est nulle : β = 0 (ou PAS = α)
- H1 : la pente de la droite de régression est différente de 0 : β ≠ 0 (ou PAS = α + β.âge)

2. Conclure à l'aide de la *P*-value : P<0.001 → P<α : rejet de H0 : acceptation de H1
La pente de la droite de régression est différente de 0 : β ≠ 0

A) Vrai
B) Faux : il s'agit de H0
C) Faux : aucun intérêt de comparer la PAS moyenne à l'âge moyen (ils sont forcément différents)
D) Faux : 0.001 est le degré de signification (P-value) du test. Le degré de signification du test est une notion distincte de l'estimation ponctuelle de la pente de la droite de régression (b = 1.0)
E) Faux

QRU 12 : A

A) Vrai
B) Faux : L'estimateur de l'ordonnée à l'origine **a** est déduit de la pente **b** et des coordonnées du point moyen (mx, my) : **a=my –bmx**
C) Faux : On note a=my – bmx, avec m**y** = mPAS = 143 et m**x** = mâge = 45
D) Faux : à 98 (a = 143 – (1 x 45) = 98)
E) Faux

J'dédicace ce DM à Pulsar, j'espère que ça t'a plu, je t'att sur le forum pour tes questions farfelues <3

Bon courage à tous, c'est la DERNIERE ligne droite donc DONNEZ TOUT, et énorme bravo à vous qui faites ce DM, ça prouve que vous bossez bien la Biostat, et c'était le bon choix à faire, vous verrez qu'elle vous le rendra ! Pleins de bisous d'amour